Portuguese-Chinese Machine Aided Translation System

Francisco Oliveira, Fai Wong, Sam Chao, Chi-Wai Tang
Faculty of Science and Technology, University of Macau
Av. Padre Tomás Pereira, Taipa, Macao
Tel: (853) 83974519, Fax: (853) 28835928
{olifran, derekfw, lidiasc, kevintang}@umac.mo

Abstract

Traditionally, all the translation work is done by individual translators separately. If they want to share translation expertise to the others, since all the translation knowledge and vocabulary are stored in their own computers, these can only be used by one-off user. Moreover, as the number of documents to be translated increases a lot nowadays, the translation task becomes impractical without the support of computer aided translation tools. As a result, the use of these tools, and their integration in the translation cycle has been one of the hottest topics in the translation community. Due to the fast development of computer technologies, networking technologies, and the fruitful results obtained from the research of Machine Translation (MT) field, computer aided translation systems are no more than that compared with the simple tools. Sophisticated translation systems may be implemented to use different MT technologies and engines, and they are able to run on different working environments and infrastructures. In this paper, a computer aided translation system based on multiple translation engines, including MT and Translation Memory, with a focus on the languages of Portuguese and Chinese is presented. Moreover, the architecture of the system is based on a Client-Server model and it has a centralized knowledge base for better management. The Portuguese-Chinese Translation (PCT) System is developed by the University of Macau.

1. Introduction

Macau is a city located at the south of China that lies at the western shore of Pearl River Delta. It is very clear that Macau presents one of the richest mixture of Western and Eastern cultures of the territory in its remarkable architecture, economic activities, and political culture. There are clear reflections and testimonies of the harmonious co-existence of both cultures in a period of interchange and assimilation extending over 450 years in the history of Macau.

Macau has currently two official languages, Chinese and Portuguese. Even after the handover in 1999, Portuguese is still considered as an official language in addition to Chinese. Although much is said that the Chinese language is becoming more important after the handover, one cannot overlook economic and diplomatic relationships that Macau has forged with other Portuguese-speaking countries with the use of the Portuguese language. Moreover, since many big enterprises in Macau and government departments are publishing their reports and announcements in both languages daily or seasonally, translation needs becomes impractical without the need of Machine Translation (MT) tools.

In the literature, different MT designs have been proposed. Rule based MT [1] approach is based on a set of linguistic grammar rules for handling the translation. Example based MT [2][3] analyzes different pieces of bilingual examples stored in parallel corpora for generating the translation. Statistic based approaches [4][5] relies on the probabilities estimated between the translation of words and the ordering of the sentences extracted from the corpora. However, each of these approaches has its strengths and weaknesses in the development of good MT systems.

In order to overcome the issues mentioned, a computer aided and hybrid translation system with a focus on Portuguese and Chinese language, PCT System, is presented in this paper. The PCT System is based on multiple translation engines, including Translation Memory; the application of Translation Corresponding Tree (TCT) [6][7] in the Example based MT paradigm as the representation structure between Portuguese and Chinese for searching and matching the fragments between bilingual texts; and the application of Constraint Synchronous Grammar (CSG) [8][9] in the Rule based MT paradigm as the language formalism for modeling the relationship between the languages simultaneously based on semantic constraints defined.

The system provides two basic translation functions. In Interactive translation, users interfere the translation process with the system sentence by sentence to evaluate the translation quality. In Automatic translation, a preliminary translation is first generated, and then users make necessary post corrections. A built-in linguistic knowledge base of more than 100,000 entries is provided, and translators can create

multiple databases easily for use in different domains. Moreover, all the translated results can be retained by the system for future use in improving the translation quality according to the experience gained.

In order to have a better management of the resources targeted to a group of translators working in the same environment, the architecture of the system is based on a Client-Server model. In the Server side, account and resources management functions, tools for evaluating the correctness of the proposed translation pairs are provided. A centralized knowledge base is resided in the server side and managed by administrators. In the Client side, translators can access and reuse the translation contribution from predecessors when they are connected to the Server. It is not only more effective and efficient but also it can increase the translation quality and reduce the cost.

This paper is organized as follows. The main functionalities of the system are presented in Section 2. The design and overall architecture are detailed in Section 3, followed by a conclusion in the last section.

2. Functionalities and Interfaces of PCT System

PCT System provides a simple interface to do the translation, which is embedded in Microsoft Word environment. All the logic and communication between PCT System and Microsoft Word are hidden from users, and they don’t need to shift between different environments to do the translation. The linguistic knowledge contains more than 100,000 entries, and PCT System provides five main functions, including the selection of the knowledge base to suit different subject needs; interactive translation with the participation of users; automatic translation to generate preliminary translations on the fly; show and hide to display the translated result against the original sentence; commit function to retain all the revised bilingual pairs to the knowledge base.

2.1 Knowledge Base Selection

According to the user’s need, they can select suitable databases to fulfill the translation requirements. Depending on the type of document to be translated, it is always useful to load technical terms and domain specific knowledge to reduce not only the translation ambiguity but also the quality. PCT System provides a simple interface to the users in switching different domain specific databases at any time.

2.2 Interactive Translation

In interactive translation, based on the selected sentences, the system first analyzes the sentences boundaries, removes all the formatting information, and for each identified sentence, they are translated one by one through the system. A translation dialog for each translated sentence to users is provided, as shown in Figure 1.

In the upper left panel, the Translation Pair shows the extracted Portuguese source sentence from the document and the Chinese translated result. At this stage, users can make any necessary amendments if necessary. In case there are any bilingual pairs similar to the extracted Portuguese sentence up to the defined threshold, they are shown in the Available MT Result, ordered from the highest to the lowest score. Furthermore, besides showing the degree of similarity of the bilingual pair extracted, extra information is given as a reference to the translators, including: citation count; the date and time of last citation; the name of the user who proposed the bilingual pair; the creation date and time; and the last modified date and time. These can further identify the bilingual pair's relevance besides considering the similarity measure estimated by the system to the users in the selection of the most suitable candidate to be replaced in the document.

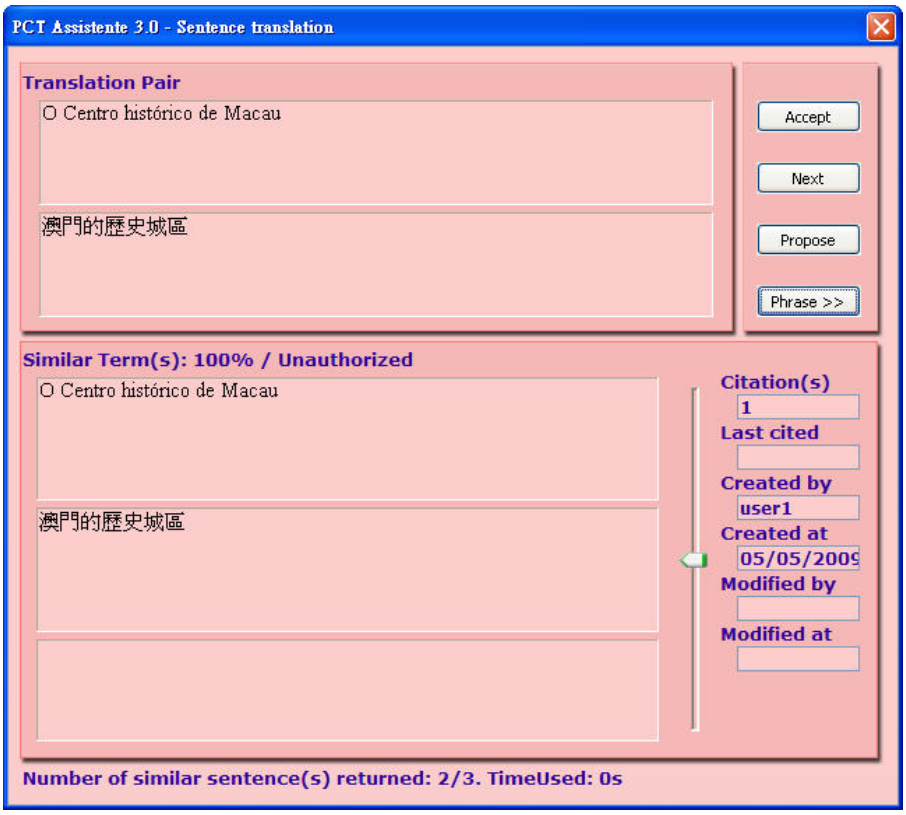

Figure 1. PCT System Translation Interface

In the upper right panel, three main functions are provided: *Accept* function takes the revised translated result out and replaces the source sentence in the document with the original formatting; *Next* function allows the user to skip the translation of the selected sentence, and no changes are made in the document; *Propose* function provides an easy way to retain the revised translation pair and propose it to the knowledge base.

PCT System provides users another way to translate a sentence in the translation dialog. Instead of generating the translation, the system allows users to select fragments of the sentence that are going to be translated. PCT System analyzes which parts of the source sentence are similar to the stored bilingual examples, and the system generates a list of found cases to the user, as shown in Figure 2. Users can make use of this list to replace any phrase that has the correct translation besides doing their own translation.

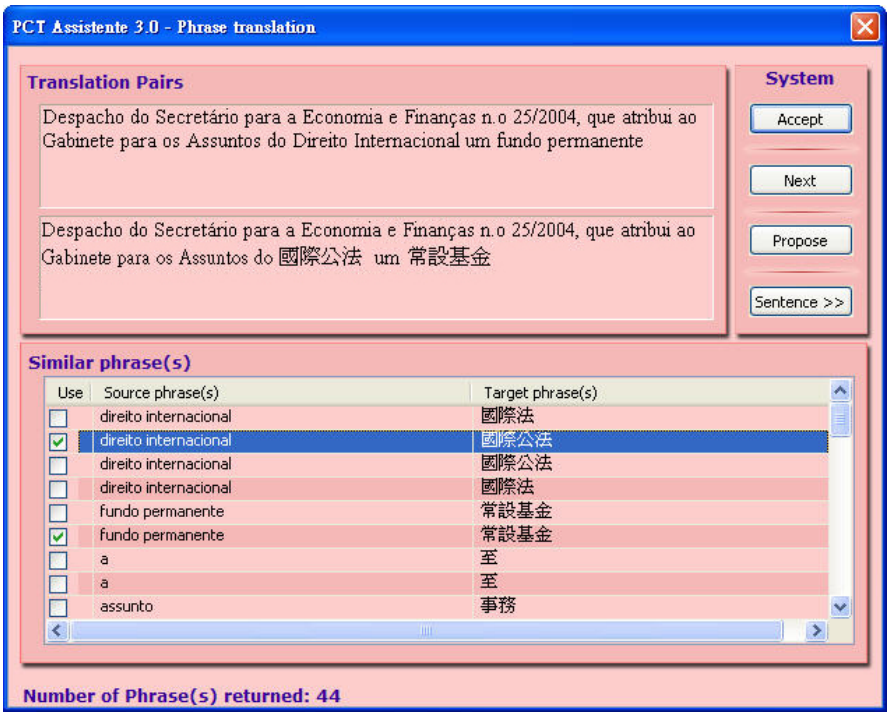

Figure 2. PCT System Phrase Translation Interface

2.3 Automatic Translation

In automatic translation, PCT System performs a preliminary translation of the whole document or selected sentences without any interaction between the system and the user. A progress bar appears and the translation that has the highest score calculated by the System for each source sentence is replaced in the document.

In some sense, this helps in making the translation work more efficient, especially

in Portuguese to Chinese translation. Since Chinese is a non-alphabetic language, and its words are based on forms of hieroglyphs, it relies on the root radicals associated to a key in the keyboard or some input methods like Pinyin. There are also Chinese handwriting pads with a stylus available in the market to help users in writing directly the character instead of typing. However, in all of the cases, users are required to be trained, and they should be very familiar in using those input methods for inserting Chinese characters in order to not degrade the effectiveness of the translation work.

As a result, by automatically generating the translation of the whole document on the fly, users could save the time in typing the translation of the whole document, and they could concentrate in post-editing the results later.

2.4 Show and Hide Bilingual Pairs

All the translation done by PCT System will replace the content of the original document. In order to let users to review the original source sentences, PCT System provides a *Show*/*Hide* function. Once enabled, as shown in Figure 3, besides the translated results, the original sentences will also appear. Each group of bilingual pair is highlighted with different colors to distinguish the boundaries. This function allows users to post-edit the translations and do not need to switch between the translated and the original document during the revision.

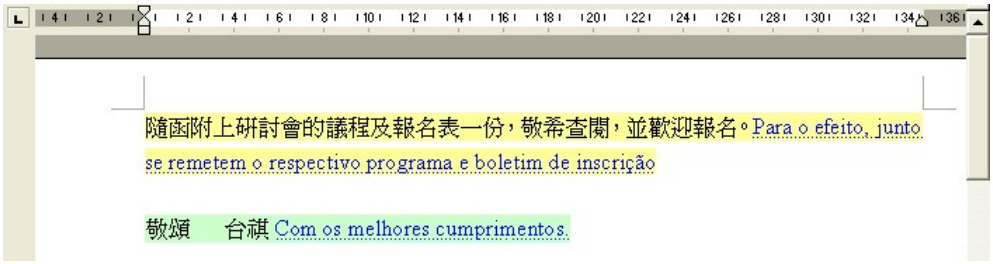

Figure 3. *Show* and *Hide* function of PCT System

2.5 Commit and Retain Bilingual Pairs

After the revision of the translated sentences, users may retain all the verified bilingual pairs into the knowledge base for future use. This functional behavior allows PCT System to increase the data in the knowledge base, and gradually, PCT System can create user's own personal translation style, better keep the consistency in the translation, and achieve more promising translation results.

3. Design of PCT System

The design of PCT System is based on a specific translation cycle that fits to the nature of translation work, as shown in Figure 4. When a document is given to the system for doing the translation, it first identifies paragraph boundaries in the extraction of candidate sentences. Next, the system will look for a set of similar example cases or fragments from the knowledge base one by one, and return to the user for references. A recalled list of examples is generated for each candidate sentence to facilitate the translation by referencing the past examples. If none of the generated results are higher than the predefined threshold, the system automatically parses and generates the translation. The suggested translation is presented to user for further revision if any amendment is necessary. This allows the user to interfere the translation process. For any new translations, if it is valuable, it can be retained and stored in the knowledge base for future use. This functional behavior allows the system to accumulate more translation knowledge during the working cycle of translators. Gradually, as the system gains more knowledge, the system can achieve more promising translation result.

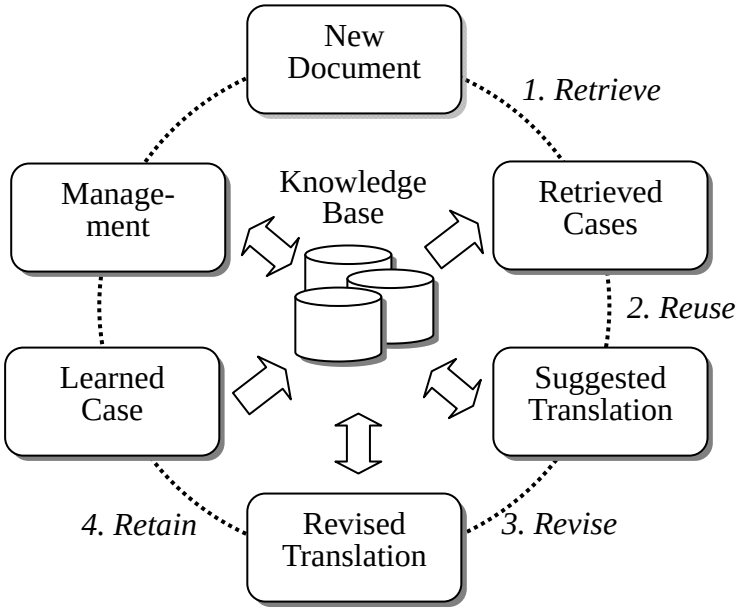

Figure 4. Translation Life Cycle

On the other hand, the architecture of PCT System is based on a Client-Server model for better resources management and sharing between a group of translators working in the same environment. Figure 5 shows the high level architecture of the system.

A centralized knowledge base is resided in the server side and managed by administrators. Moreover, in the Server side, the system provides 1) management functions for monitoring account information and status; 2) tools for evaluating the correctness of the translation pairs proposed by the translators. In the Client side, once connected to the system, translators can access and reuse the translation pairs stored in the centralized knowledge base as well as the previous contributions made by predecessors. Moreover, the system provides translation functions as well as a local knowledge base for users to store their translation knowledge if they would like to keep their information private from others.

This architecture is very effective in terms of resource management, since all the knowledge and vocabulary are centralized, they can be uniquely managed and shared through the network. Moreover, the translation quality and output can be kept more coincident with high efficiency done by different translators. The cost in maintaining the system can be greatly reduce and it can form the basis for constructing a Machine Translation industry.

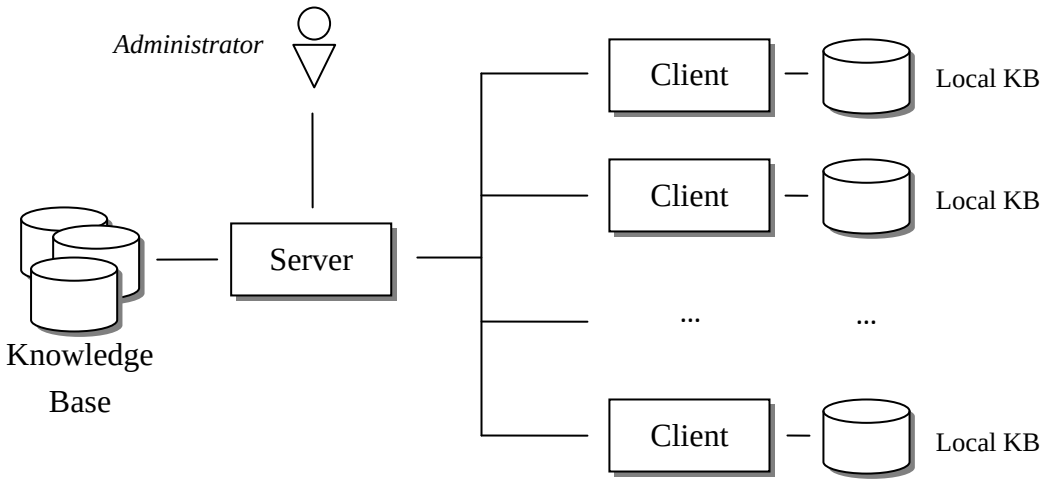

Figure 5. High Level Architecture of PCT System

4. Conclusion

PCT System is developed to fulfill the high demand of translation work in Macau. The main objective of PCT System is to fulfill the high demand of translation work in Macau as well as to increase the translation efficiency in the daily work of the translators. A simple interface is provided to the translators and all the underlying technologies, are transparent to them. In other words, translators just need to use their familiar Word processing environment to do their translation work, and all the

translation tasks are accomplished automatically by the PCT System through the network connection. Currently, PCT System has been distributed to different Macau government departments, schools, universities, local firms, etc, and their feedback fully reflect the merits of our unique research findings in the city.

Acknowledgement

This work was partially supported by the Research Committee of University of Macau under grant UL019/09-Y2/EEE/LYP01/FST, and also supported by Science and Technology Development Fund of Macau under grant 057/2009/A2.

References

1. Bennett, W.S., Slocum, J.: The LRC Machine Translation System. Computational Linguistics 11(2-3), 111--121 (1985)
2. Brown, R.D.: Example-Based machine translation in the Pangloss system. In: Proceedings of the 16th International Conference on Computational Linguistics (COLING-96), pp. 169--174. Copenhagen, Denmark (1996)
3. McTait, K.: Translation Pattern Extraction and Recombination for Example-Based Machine Translation. PhD Thesis, Centre for Computational Linguistics, Department of Language Engineering, UMIST (2001)
4. Brown, P.F., Cocke, J., Della Pietra, S.A., Della Pietra, V.J., Jelinek, F., Lafferty, J.D., Mercer, R.L., Roossin, P.S.: A Statistical Approach to Machine Translation. Computational Linguistics 16(2), 79--85 (1990)
5. Lopez, A.: Statistical Machine Translation. ACM Computing Surveys, Vol. 40, No.3, Article 8. (2008)
6. Wong, F., Hu, D. C., Mao, Y. H.: A Flexible Example Annotation Schema: Translation Corresponding Tree Representation. In: Proceedings of the 20th International Conference on Computational Linguistics, pp.1079--1085. Switzerland, Geneva (2004)
7. Tang, C. W., Wong, F., Leong, K. S.: Application of Translation Corresponding Tree (TCT) Annotation Schema for Chinese to Portuguese Machine Translation. In: Proceedings of the Tenth International Conference on Enhancement and Promotion of Computational Methods in Engineering and Science (EPMESC-X), pp.1105--1109. Sanya (2006)
8. Wong F., Hu D.C., Mao Y.H., Dong M.C., Li Y.P.: Machine Translation Based on Constraint-Based Synchronous Grammar. In: Proceedings of the 2nd International Joint Conference on Natural Language (IJCNLP-05), pp. 612--623. Jeju Island, Republic of Korea (2005)
9. Oliveira, F., Wong, F., Leong, K.S., Tong, C.K., Dong, M.C.: Query Translation for Cross-Language Information Retrieval by Parsing Constraint Synchronous Grammar. In: Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, pp. 4003-4008. Hong Kong (2007)